Anais do 6º Congresso Brasileiro de Agroinformática – SBIAgro 2007, São Pedro, Brasil, 8-11 outubro 2007, Embrapa Informática Agropecuária, p.41-45.

USANDO O SOFTWARE GALICIA PARA CALCULAR SIMILARIDADE ENTRE ONTOLOGIAS DO DOMÍNIO AGROPECUÁRIO

KLEBER XAVIER SAMPAIO DE SOUZA[1]
SILVIA MARIA FONSECA SILVEIRA MASSRUHÁ[2]
PAULO ROBERTO BATISTA JUNIOR[3]

RESUMO: Ontologias têm sido criadas para numerosos domínios do conhecimento, tais como vinhos, partes de automóveis, processos biológicos, metadados de informação geográfica, processos agrícolas etc. Ao mesmo tempo que esta abordagem de criação descentralizada e desordenada de ontologias é altamente desejável, ela dá origem a um outro problema: como comparar ontologias de domínios correlatos de forma que o conhecimento possa ser reusado. Neste artigo, apresenta-se uma metodologia de utilização do software Galicia para computar a similaridade entre conceitos de ontologias distintas. Esta medida de similaridade foi construída com base na Análise Formal de Conceitos, uma técnica que aplica o formalismo matemático desenvolvido por Evariste Galois para a análise de dados e identificação da estrutura de conjuntos interrelacionados.

PALAVRAS-CHAVE: ontologia, reticulados de galois, análise formal de conceitos, tesauros, recuperação de informação.

USING GALICIA SOFTWARE TO CALCULATE SIMILARITIES AMONG ONTOLOGIES IN AGRICULTURAL DOMAIN

ABSTRACT: Ontologies have been created for many knowledge domains, such as wine, car parts, biological processes, geographic information metadata, agricultural processes etc. However, whereas this decentralized and unordered approach is highly desirable, it gives rise to other problems, such as how to compare ontologies in correlated domains in such a way that the knowledge can effectively be reused. This article discusses how Galicia software can be used to compute the similarity among concepts from different ontologies. This similarity measure was constructed taking Formal Concept Analysis (FCA) as a basis. FCA is a technique that applies the mathematical formalism developed by Evariste Galois to analyze datasets and identify the structure of inter-related sets.

KEY-WORDS:ontologies, galois lattices, formal concept analysis, thesauri, information retrieval.

1. INTRODUÇÃO

Ontologias, ou especificações explícitas de uma conceituação (GRUBER, 1993), têm sido criadas para um grande número de domínios de aplicação com o intuito principal de permitir o processamento automático de informação por agentes de software que, automática e autonomamente, seriam capazes de processá-las e identificar o significado preciso dos termos e relações nelas contidos.

Ontologias são constituídas de símbolos (termos do domínio de aplicação) e relações entre símbolos. Quando confrontados com duas ontologias de domínios correlatos, porém distintos,

[1] Doutor em Telemática, Embrapa Informática Agropecuária, kleber@cnptia.embrapa.br
[2] Doutora em Computação Aplicada, Embrapa Informática Agropecuária, silvia@cnptia.embrapa.br
[3] Graduando em Engenharia de Computação, Unicamp, Embrapa Informática Agropecuária, Bolsista CNPq de Iniciação Científica, paulorbj@cnptia.embrapa.br

como **cultivo do feijão** e **cultivo do milho**, por exemplo, que possuem uma série de termos aparentemente iguais mas posicionados em pontos diferentes, os agentes de software não são capazes de afirmar se se tratam de conceitos iguais, ou apenas uma coincidência do ponto de vista simbólico.

Quando lidam com ontologias distintas, os agentes precisam realizar um emparelhamento de seus símbolos, considerando, naturalmente as relações entre os símbolos. Este emparelhamento recebe o nome de **alinhamento de ontologias** (SOUZA & DAVIS, 2004).

Neste artigo, propõe-se a utilização do software Galícia, desenvolvido na Universidade de Montreal[4], na determinação do cálculo de similaridade entre os termos de duas ontologias. O foco do trabalho é na automatização deste cálculo, e não no desenvolvimento da medida de similaridade em si. Esta foi desenvolvida em outras publicações dos autores e requer base matemática avançada (SOUZA et al. 2006). Nessas publicações, contudo, os detalhes da utilização do Galícia não foram abordados.

2. MATERIAL E MÉTODOS

O primeiro passo necessário ao alinhamento de duas ou mais ontologias é o estabelecimento de uma base comum de comparação. Na metodologia desenvolvida em SOUZA et al. (2006), esta base é formada por um **reticulado de galois**. Este reticulado é gerado utilizando-se a Análise Formal de Conceitos (FCA[5]) (GANTER & WILLE, 1999; WILLE, 1982). Basicamente, o que a FCA nada mais é do que uma representação espacial da relação existente entre os vários subconjuntos de um conjunto. A Figura 1 ilustra parte dos objetos das ontologias de gado de corte e gado de leite. A FCA parte de uma matriz como esta, contendo objetos nas linhas e atributos nas colunas e determina quais subconjuntos de objetos estão relacionados com quais subconjuntos de atributos.

Figura 1. Contexto formal contendo as ontologias de gado de corte (A) e gado de leite (B)

[4]Galicia: Galois Lattice Interactive Constructor. http://www.iro.umontreal.ca/~galicia/
[5]Formal Concept Analysis (FCA)

Como se necessita de um conjunto de atributos para realizar a comparação entre as ontologias, utilizou-se, para este fim, termos do tesauro Agrovoc. Originalmente, cada uma das ontologias possui seus nós catalogados com base neste tesauro. Os objetos da ontologia de gado de corte iniciam-se com a letra A e a de gado de leite, com a B. Tendo-se a matriz de objetos e atributos, pode-se agora aplicar a FCA obtendo-se o reticulado ilustrado na Figura 2. Nessa figura cada nó do reticulado representa um conceito, contendo objetos e atributos. Para exemplificar, os objetos **X** e **Y** possuem como atributos **a**, **b** e **c**. Contudo, **X** possui **a** e **b** como atributos e **Y** possui **a** e **c** como atributos. Isto significa que **X** e **Y** possuem **a** como atributo comum.

Na medida de similaridade desenvolvida em SOUZA et al. (2006), apenas os elementos estruturantes com um único nó pai foram considerados. Os nós com rótulos **a**, **b** e **c** na Figura 2 são estruturantes, enquanto que os com rótulos **X** e **Y** não o são. Note-se, por exemplo, que **X** possui **a** e **b** como pais. Nesta medida, conta-se o número de elementos estruturantes **(estX e Y)** que **X** e **Y** possuem em comum, dentre todos os caminhos entre eles e o nó raiz; o número de elementos estruturantes que apenas **X** possui, ponderado por um fator de 0.5; e os estruturantes ligados apenas a **Y,** também ponderados por 0.5.

$$\mathrm{Sim}(X,Y) = (estX \ \ e \ \ Y) \ / \ ((estX \ \ e \ \ Y)+0.5(estX)+0.5(estY))$$

Aplicando-se esta fórmula para X e Y da Figura 2, ter-se-ia como resultado:

$$\mathrm{Sim}(X,Y) = (1) \ / \ ((1)+0.5(1)+0.5(1)) = 0.5$$

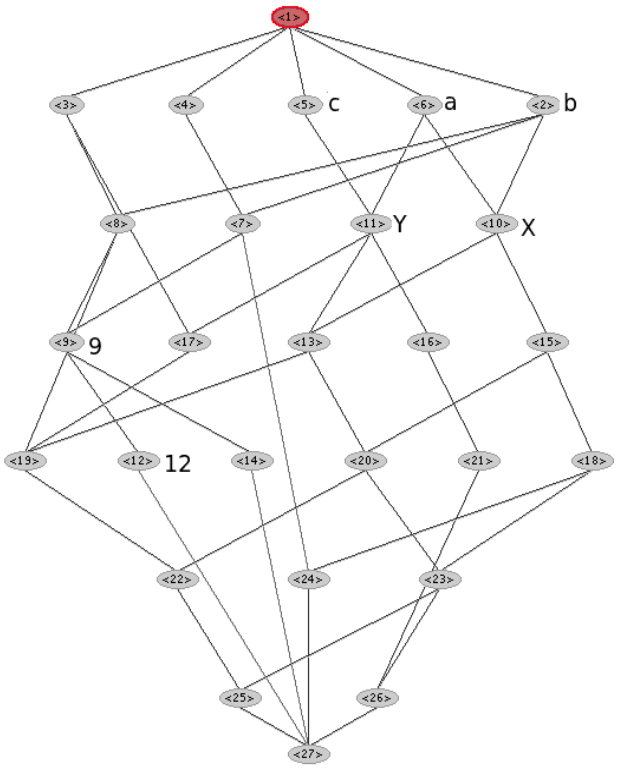

Figura 2. Reticulado de Galois correspondente ao contexto da Figura 1

No exemplo da Figura 2, **X** e **Y** têm em comum o nó **a**, este nó é estruturante por introduzir o atributo **IntensiveHusbandry** (abreviado para IntensiveH), e as suas diferenças estão nos nós **c**, que introduz **BeefCattle**, e **b**, que introduz **FeedingSystems**.

3. RESULTADOS E DISCUSSÃO

Como o desejado é a automatização deste cálculo, que permite a determinação da medida de similaridade entre todos os objetos do contexto da Figura 1, acoplou-se a saída da ferramenta Galícia com um programa Java® que realiza o cálculo propriamente dito. A Figura 3 ilustra este processo. O acoplamento entre o Galicia e o programa que realiza o cálculo de similaridade é feito por meio de um XML[6] gerado pelo próprio Galicia.
Visualmente, a avaliação de similaridade entre conceitos é muito fácil para um ser humano quando o número de nós do reticulado é pequeno. Entretanto, à medida em que o número de nós cresce, o grafo que representa o reticulado torna-se muito enovelado, dificultando muito esta avaliação. De qualquer maneira, a avaliação do ponto de vista numérico, dada pela medida de similaridade, resulta em um conjunto de valores que podem ser utilizados tanto por avaliadores humanos quanto por programas (agentes de software).

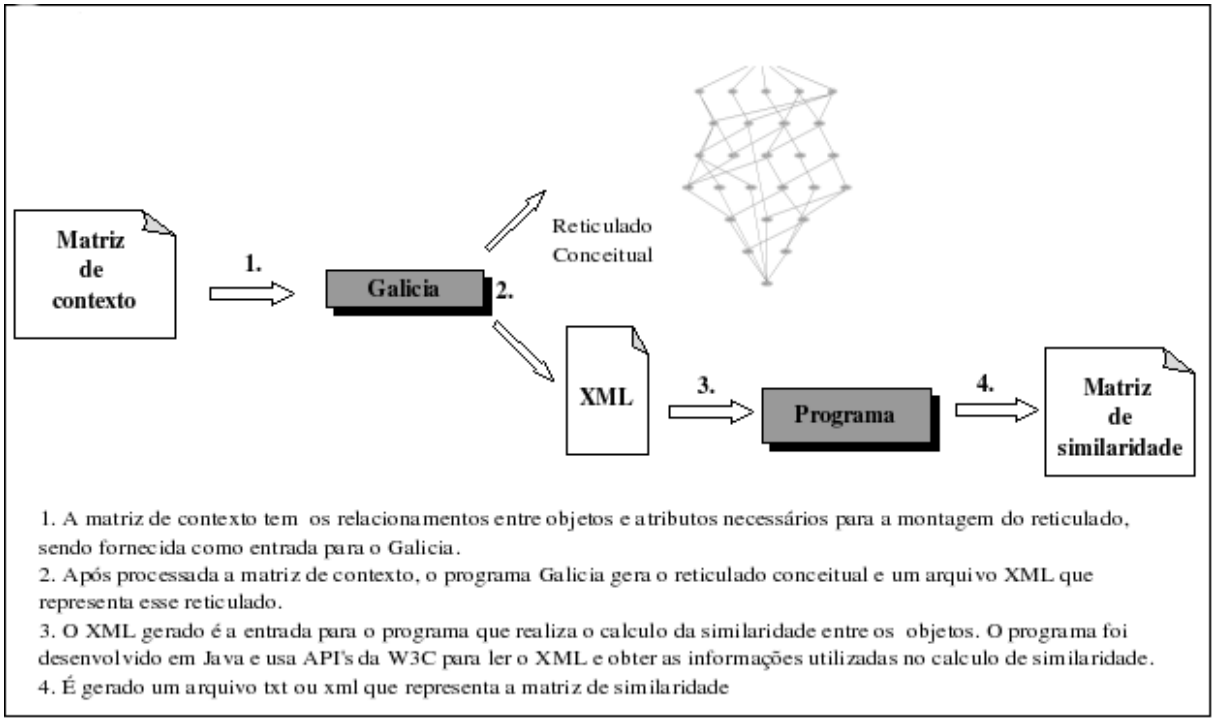

Figura 3. Processo de geração da matriz de similaridade

A Tabela 1 mostra parte da tabela de similaridades entre as ontologias das Figuras 1 e 2. A matriz completa para o exemplo é uma matriz de 27 X 27, da qual tomou-se para fins de ilustração apenas a submatriz de 12 X 12, correspondente aos 12 primeiros nós comparados com os demais de 1 a 12. Por exemplo, a similaridade entre os nos identificados como X e Y na Figura 2, ou seja nós 10 e 11, é 0,5. Já entre os nós 9 e 12, a similaridade é 0,86, que é bastante alta, dado que o nó 9 compartilha a maioria dos elementos estruturantes do nó 12.

[6]XML – eXtented Markup Language. Linguagem desenvolvida pelo World Wide Web Consortium (W3C) para facilitar troca de dados entre sistemas.

Tabela 1. Processo de geração da matriz de similaridade

Nó	1	2	3	4	5	6	7	8	9	10	11	12
1	1	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0,67	0,67	0,5	0,67	0	0,4
3	0	0	1	0	0	0	0	0,67	0,5	0	0	0,4
4	0	0	0	1	0	0	0,67	0	0,5	0	0	0,4
5	0	0	0	0	1	0	0	0	0	0	0,67	0
6	0	0	0	0	0	1	0	0	0	0,67	0,67	0
7	0	0,67	0	0,67	0	0	1	0,5	0,8	0,5	0	0,67
8	0	0,67	0,67	0	0	0	0,5	1	0,8	0,5	0	0,67
9	0	0,5	0,5	0,5	0	0	0,8	0,8	1	0,4	0	0,86
10	0	0,67	0	0	0	0,67	0,5	0,5	0,4	1	0,5	0,33
11	0	0	0	0	0,67	0,67	0	0	0	0,5	1	0
12	0	0,4	0,4	0,4	0	0	0,67	0,67	0,86	0,33	0	1

4. CONCLUSÃO

O software Galicia mostrou-se perfeitamente adequado ao cálculo de medidas de similaridade que considerem a estrutura do reticulado que o contexto representa. Isto ocorre porque o Galicia possui como saída o reticulado expresso em XML, facilitando o acoplamento de desta com aplicações específicas para o cálculo de similaridade.

5. REFERÊNCIAS

GANTER, B.; WILLE, R. **Formal Concept Analysis: Mathematical Foundations**. Springer, Berlin-Hidelberg-New York, 1999.

GRUBER, T.R.. A Translation Approach to Portable Ontologies Specifications. **Knowledge Acquisition**, v. 4, pp 199-220, 1993.

SOUZA, K.X.S.; DAVIS, J. Aligning Ontologies and Evaluating Concept Similarities. **Lecture Notes in Computer Science (LNCS)**, v. 3291, pp 1012-1029, 2004.

SOUZA, K.X.S.; DAVIS, J.; EVANGELISTA, S.R.M.; Aligning Ontologies, Evaluating Concept Similarities and Visualizing Results. **Journal on Data Semantics V**, LNCS 3870, pp. 211-236, 2006

WILLE, R. Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. In RIVAL, I. (ed.) Ordered Sets, Volume 83 of NATO Advanced Study Institute Series C. Reidel, Dordrecht, p. 445-470, 1982.